TDT 2003

Preliminary Call for Participation

http://www.nist.gov/TDT

You are invited to participate in the Topic Detection and Tracking (TDT) evaluation workshop for 2003. TDT research aims to develop algorithms for the automatic organization of news stories by the real-world events that they describe. The source media for the workshop includes newswire text as well as the audio from radio and television broadcasts, all in English, Chinese, and Arabic. Although it is possible to work in the original media and language, most TDT researchers take advantage of provided translations and/or speech recognition results.

TDT 2003 is the sixth in a series of open evaluations that have investigated several aspects of this problem. Researchers who are interested in the following topics are encouraged to find out more about TDT 2003: high accuracy retrieval of documents, text filtering, cross-language issues, machine translation, speech recognition, text segmentation, compensating for degraded quality text, novelty detection, or other similar topics.

TDT evaluation tasks for 2003 are currently expected to be:

- New event detection
- Story link detection
- Segmentation
- Topic detection
- Tracking

Other tasks are being considered and the final evaluation specifications are being debated.

For more information, please contact James Allan (allan@cs.umass.edu) or Jonathan Fiscus (jonathan.fiscus@nist.gov) to be added to the mailing lists. Or send "subscribe tdt-distrib" to majordomo@ldc.upenn.edu to be added to the mailing list.

Important dates
(all in 2003)

Sept 2	Eval data released
Oct 3	Results due at NIST
Nov 17-18	TDT 2003 workshop

(Dates are subject to change)

TDT Corpora

The **TDT-2 corpus** includes 80,000 stories from the first six months of 1998. The stories come from six English sources and three Chinese sources. TDT-2 also includes 100 event-based topics that have been judged for relevance in both languages.

The **TDT-3 corpus** is made up of 40,000 stories from the last three months of 1998 as well as 120 topics judged for relevance against those stories. The stories come from eight English, three Chinese, and four Arabic sources. (Only sixty of the topics are judged against the Arabic stories).

The **TDT-4 corpus** was developed for the TDT 2002 evaluation. The text corpus will be reused, and an additional set of topics will be annotated for the 2003 evaluation. It includes 45,000 stories from the end of 2000, from eight English, seven Chinese, and four Arabic sources.